Evolutionary Computation and the C-value Paradox

Sean Luke
George Mason University
4400 University Drive MS# 4A5
Fairfax, VA 22030 USA
sean@cs.gmu.edu

ABSTRACT

The C-value Paradox is the name given in biology to the wide variance in and often very large amount of DNA in eukaryotic genomes and the poor correlation between DNA length and perceived organism complexity. Several hypotheses exist which purport to explain the Paradox. Surprisingly there is a related phenomenon in evolutionary computation, known as code bloat, for which a different set of hypotheses has arisen. This paper describes a new hypothesis for the C-value Paradox derived from models of code bloat. The new explanation is that there is a selective bias in preference of genetic events which increase DNA material over those which decrease it. The paper suggests one possible concrete mechanism by which this may occur: deleting strands of DNA is more likely to damage genomic material than migrating or copying strands. The paper also discusses other hypotheses in biology and in evolutionary computation, and provides a simulation example as a proof of concept.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences-Biology and genetics

General Terms

Experimentation

Keywords

C-value Paradox, Evolutionary Genetics, Code Bloat, Genetic Programming, Theoretical Biology

1. INTRODUCTION

Evolutionary computation derives much of its inspiration, and indeed its very name, from evolution and genetics. Ideas and terminology in biology are readily adapted and repurposed, sometimes inappropriately, to form new approaches in stochastic optimization and search. This paper instead goes the other way. It suggests one possible solution to the biology's C-value Paradox inspired by explanations of similar phenomena in evolutionary computation.

The *C-value* of an organism is the amount of DNA in the organism's genome. The variation in DNA size across eukaryote species is nothing short of astonishing. According to Gregory [11], the genome of *Amoeba dubia* is over 200,000 times the size of the genome of *Encephalitozoon cuniculi* (and 200 times the size of the *Homo sapiens* genome). Surprisingly, there is very little variation among prokaryote genomes,[1] and generally speaking prokaryote genomes are very small relative to those of most eukaryotes.

In 1971, Thomas [23] coined the term *C-value Paradox*[2] to denote the unexpected lack of relationship between the presumed complexity of an organism and its C-value (consider the C-value of the *Amoeba*). Notably, there often exists wide variation in DNA content among closely related species. This was seen as a paradox because at the time genomes were believed to be simply sets of genes. Why then, it was asked, would DNA be so large if the number of genes was expected to be small?

As was later discovered, much of a genome's DNA may not code for any gene at all. This apparent non-coding DNA was termed, perhaps inappropriately, "junk DNA".[3] The existence of non-coding DNA provided one answer to the question of how it was *possible* for DNA to be large relative to organismal complexity or its number of genes, but not the *reason* why it was large. As such the C-value Paradox survives today partly in this form: why is there so much non-coding DNA in some organisms? This is the question which this paper will address. Other major questions (not discussed in this paper) include: why is there so much variation in amount of non-coding DNA across species? And: why do some "complex" organisms have few genes while some "simple" organisms have many?

[1]For the forgetful: a prokaryote lacks a nucleus and various other organelles, and its genome is a single loop of DNA which wends its way through the cell. A eukaryote has organelles and a nucleus, and a genome consisting of one or more chromosomes, plus certain auxiliary DNA such as mitochondrial DNA.

[2]As the C-value Paradox is not a "paradox" any more per se, Gregory [11] has suggested using the term *C-value Enigma* instead: but I believe the earlier term is still in common use.

[3]Since then some junk DNA has been found to serve one purpose or another in the genome, so at least *part* of it may not be junk after all. At any rate, the term "junk" is falling out of favor.

GECCO'05, June 25–29, 2005, Washington, DC, USA.

Eukaryotic non-coding DNA is considerable. In *Homo sapiens* extragenic DNA accounts for over 70% of the genome. Within the boundaries of a gene, on average 90% of the material consists of introns and other non-coding material. Only 3% of the human genome actually codes for protein, and some percentage of those genes may have no real purpose [4]. Interestingly, prokaryotic DNA tends to have very little "junk", though this may be explained by selection pressuring prokaryotes to be small and to reproduce rapidly (small DNA size is correlated with both of these features). Many large, multicellular eukaryotes do not have this constraint, though this does not explain why some small, single-celled eukaryotes (such as *Amoeba*) have among the largest genomes in existence.

There are various hypotheses which attempt to solve the C-value Paradox and explain why many genomes may consist of so much non-coding DNA. Each has advantages and difficulties. In this paper I will enumerate these methods, and then discuss the curiously related "bloat" phenomenon found in various evolutionary computation and related stochastic optimization techniques (particularly genetic programming). Following this, I will propose a new hypothesis which attempts to explain the C-value Paradox, and which is interestingly derived not from biological foundations but from inspiration drawn from the results of these optimization techniques. Even more surprisingly, to my knowledge this hypothesis has not been proposed in the biological literature, despite its relative obviousness. The new hypothesis is not intended to replace the others: indeed I believe that several of them are likely to be true. Instead, I think that several forces are behind the C-value Paradox, and that the new hypothesis may be one of them.

2. GENOME GROWTH HYPOTHESES

Most explanations for the C-value Paradox rely on some underlying force involved in producing additional DNA independent of the selection process. Acted on by this force, the genome will grow in size until the length of the genome places the organism at selective disadvantage [17]. One source of this selective disadvantage is that DNA size is correlated with cell and nucleus size and with (slower) cell division rate. If, for example, the organism required rapid cell division, then a large genome would be selected against. Doolittle and Sapienza also note that underlying genome-growth events (they referred to transposition as described below) might also be "frankly destructive" to the genome, which in turn could produce selection against them [10].

What forces might act upon the genome to cause it to grow? There are several possibilities, including:

- *Bulk Modifications* might erroneously occur during DNA replication.
- *Strand Slippage* is the unintended repetition of a few DNA base pairs during DNA replication, possibly caused by incorrect DNA repair. Charlesworth *et al* state that "*In vitro* studies suggest that strand slippage during DNA replication is a major cause of the observed length polymorphism of microsatellites[4] within populations." [7]
- *Transposable Elements* are sequences of DNA capable of transposing to another location in the DNA strand. These may be divided into two classes: elements which act on the DNA directly from the DNA itself, and "retroelements" which insert in the DNA from RNA (for example, retroviruses) [7]. Transposable elements may simply move themselves to a new location in the DNA; or they may insert *copies* of themselves into the DNA, resulting in genome growth. Again, Charlesworth *et al* state that "much of the moderately dispersed repeated DNA of eukaryotes appears to consist of transposable elements...." [7]

  Copying via transposition events is also likely the primary way by which new genes are formed. If an existing gene is important to the function of the organism, then mutating that gene (to convert it to a new gene) is likely to be deleterious. But if the gene were duplicated through the copying of a DNA sequence, then the original gene would be free to mutate as there is now another copy of the gene producing the original gene's crucial RNA [16].

Gregory [11] describes four prominent hypotheses for genome growth: Junk DNA, Selfish DNA, Nucleoskeletal, and Nucleotypic, and provides an excellent comparison of them. I list these four plus two more.

The Junk DNA Hypothesis. This hypothesis argues that transposable elements, strand slippage, and other gene-accumulating events might simply accumulate on their own independent of selection. These changes then are spread through the population via genetic drift [11, 18]. Dawkins sums it up thus: "The simplest way to explain the surplus DNA is to suppose that it is a parasite, or at best a harmless but useless passenger, hitching a ride in the survival machines created by the other DNA." ([8], p. 42)[5]

The Selfish DNA Hypothesis. This hypothesis may be viewed as an extension of the Junk DNA Hypothesis that asserts an actual selective force which drives the transposable element-copying procedure [10, 17]. Here transposable elements will repeatedly copy themselves elsewhere into the genome, and those copies will copy themselves, etc. resulting in considerable genome growth.

The "selective" force is not at the organismal level but *within the genome*. Doolittle and Sapienza argue: "transposability itself ensures the survival of the transposed element..." [10]. Thus because it is *capable* of transposing, a transposable element is more likely to survive deleterious mutation than a non-transposable element. This is plausible, though given the relative rarity of a element-damaging event when compared to the death or survival of an organism, such a force seems very weak compared to the selective force on genes to ensure survival of their host organism. Additionally, while the Junk DNA hypothesis might predict a linear increase in genome size, Selfish DNA suggests an exponential increase in size as children of selfish copiers themselves begin copying until the length becomes a hindrance to fitness. This seems problematic.

The hypothesis's name is derived from from Dawkins' *The Selfish Gene* [8], which posits that selection must be viewed

[4]Short, heavily repeated sequences of non-coding DNA are termed satellites, minisatellites, or microsatellites, depending on the length of the sequence.

[5]This statement of Dawkins's may also be used to argue partly for the Selfish DNA hypothesis.

at the gene level rather than at the genome or organismal level. The borrowing is unfortunate, as Dawkins' treatise deals with the spread of genes through populations, not genomes [11], and more importantly, his "selfish genes" are selected for because they have an *effect on their organism*, whereas Orgel and Crick [17] define selfish DNA as having the following features: "(1) It arises when a DNA sequence spreads by forming additional copies of itself withing the genome. (2) It makes *no specific contribution to the phenotype*."[6]

It is important to note that both of these hypotheses are additive only: they presume that the mechanisms for increasing genome size are much more prevalent than those for reducing it [11]. Instead they rely on selection as the force for maintaining a cap on unrestricted genome growth. However, it is known that genome-reduction mechanisms do exist [6, 11]. Indeed, such mechanisms have had a dramatic effect in reducing genome size in certain organisms such as *Arabidopsis* [9].

The Nucleoskeletal and Nucleotypic Hypotheses. These hypotheses rely on the positive relationship between the size of the genome and cellular features. The nucleoskeletal theory argues that organisms have an optimal cell size for maintaining metabolism, cell division, etc. [5] Selection for this cell size in turn places selective pressure on producing an appropriate nucleus size relative to the cell size, and this in turn selects for some mechanism (ostensibly an appropriately sized genome) for producing the desired nucleus size. The nucleotypic hypothesis is closely related, but instead asserts a more direct relationship between genome size and resulting cell size and other cellular parameters. Selection for these parameters in turn puts pressure on an appropriate genome size [2]. For extensive elaboration on these two hypotheses, see [11].

Unfortunately, neither of these hypotheses suggests a function by which genome growth occurs: they simply attribute it to selection for those external features which genome size can affect. No real causal link asserted. While a larger genome may be able to produce a larger nucleus, a larger cell, and changes in cellular parameters, surely these features can be come by through other means, such as support mechanisms coded in the organism's DNA. The complex task of increasing or decreasing DNA to produce the desired effect seems to be the most roundabout of several approaches to meeting the selective needs of the cell.

The Defense Against Recombination Hypothesis. This is the term I give to the hypothesis casually suggested by Pagel [18].[7] The idea is that large swaths of non-coding DNA may be selected for because in some way they protect against the possibility of recombination during the cell replication process, perhaps by encouraging recombination events to occur in the non-coding regions of DNA rather than the coding regions. As informal evidence Pagel points to mitochondrial DNA, which does not undergo recombination and also is believed to lack non-coding DNA.

I feel that one difficulty with this hypothesis is that it presumes that the presence of one recombinative event somehow diminishes the likelihood of a related nearby event. Were this the case, then non-coding DNA, and particularly such DNA designed to encourage recombination, could thus somehow stifle recombination in more crucial regions. However I am not aware of support for this presumption.

The Genome Selective Benefit Hypothesis. There exist examples in the literature where non-coding DNA act sufficiently on the genome itself as to possibly impart a selective benefit. For example, intron size has been shown to be both positively correlated and negatively correlated with recombination rate, depending on the species [19]. Whether there is a general selective benefit remains to be seen: but it remains worthwhile mentioning this as a possible hypothesis.

3. CODE BLOAT IN EVOLUTIONARY COMPUTATION

One of the unexpected, coincidental consequences of evolutionary computation deriving its inspiration from evolution and genetics is that the EC field too has its own C-value Paradox of a sort. This is the phenomenon of *code bloat* (or as Bill Langdon calls it, "survival of the fattest"). Bloat is the tendency for individuals' genomes to grow significantly in size during an evolutionary run, and to do so in a fashion unrelated to significant improvement in fitness. The phenomenon is prevalent in genetic programming, though it has cropped up in a variety of evolutionary computation contexts — in fact, the first reported example of bloat occurred in Pitt-approach rule programs [20]. While we are beginning to better understand the underlying causes of code bloat, at least in tree-based genetic programming representations, we are still a ways off from constructing methods to counter it: instead we tend to resort to the blunt instrument of parsimony pressure.

For further information on bloat, theoretical models of the phenomenon, methods for countering it, and examples of bloating features, see [13]. Here is a summary of the major theories, all of which primarily concern themselves with evolutionary computation genomic representations in the forms of strings or trees.

Hitchhiking. This model places blame on the presence of *introns*, regions of code which serve no purpose in the genome. The idea is that such chunks (the "hitchhikers") attach themselves to parents of "important" active code in such a way that when the active code is propagated from individual to individual, the hitchhikers come along for the ride [22]. This hypothesis has many parallels to the Junk DNA hypothesis, and in fact Dawkins's quote refers to "hitchhikers".

Defense Against Crossover. Here blame is placed on a subset of introns called *inviable code*. These are regions of code which serve no purpose, and furthermore cannot be

[6]Dawkins himself and Stephen J. Gould openly disagree on whether Selfish DNA has much to do with the Selfish Gene (each finding the others' reasoning "wrong but interesting"). Gould argues that "selfish genes increase in frequency because they have effects on bodies...selfish DNA increases in frequency...because it has no effect on bodies...." whereas Dawkins, enamored with Orgel and Crick's borrowing of selfish genes to make the selfish DNA argument, puts forth various — and in my view unsatisfactory — responses ([8], p. 275).

[7]I also note with interest that Pagel's suggestion predates genetic programming's defense against crossover theory by two years.

replaced with any code which can possibly serve a purpose. Modification (via crossover or other recombination; or some form of mutation) of these regions cannot change the fitness of the individual in any way. Late in the evolutionary process, individuals in the population have largely converged in fitness, and so any major fitness change is likely downward. Thus it is in the individual's interest to not change at all. As the breeding step in genetic programming consists of a single recombinative event, selection can thus pressure the individual to sprout large amounts of inviable code to reduce the probability that viable parts of the genome will be modified (and likely damaged). Defense against crossover has long been a popular model ([3, 15, 1] among many others).

Pagel's C-value Paradox hypothesis, which I termed "defense against recombination", is tantalizingly similar to defense against crossover as described here. But keep in mind that DNA copying not like the artificial breeding mechanism in genetic programming: in real DNA it is not clear that the probability of recombination in one DNA region has a strong effect on the probability of recombination elsewhere. However in GP the relationship is very strong indeed: if an event occurs in one place it *will not* occur elsewhere.

Removal Bias. This GP tree-specific model also places blame on inviable code. Here inviable code takes the form of subtrees typically located in the fringes of the genetic programming tree genome [21]. If modification occurs in inviable code regions, the removed subtree must therefore generally be small compared to modification of viable code regions (further up in the tree). Thus removing a small subtree is more likely to have no effect on the individual. But there is no such bias for *inserting* subtrees in these regions: a subtree of any size can be added to an inviable code region with no effect. Large removed subtrees are selected against, but not large inserted subtrees.

Fitness Causes Bloat. This somewhat abstract hypothesis argues that there are many more highly-fit large genomes than highly-fit small genomes, if only because there are more large genomes than small ones in general. But genetic programming typically starts with very small genomes to begin with, and so bloat might simply be the system moving towards equilibrium [12].

Modification Point Depth. This GP tree-specific hypothesis argues that there is a correlation between the depth of the modification point in a tree genome and the likely *survivability* of the child (defined as how well the child performs against its peers in the population — such as how often the child or later ancestors are selected for reproduction). Deeper modification points are correlated with higher survivability. But the choice of deeper modification points also is correlated with larger parents (who generally create larger children), and deeper modification points tend to root smaller subtrees, creating a removal bias somewhat similar to that discussed earlier. But unlike the defense against crossover and removal bias models, modification point depth does not rely on viable vs. inviable code: it posits a gradient of viability over various modification points [13]. Indeed, the removal bias model may be viewed as a specific subset of the modification point depth hypothesis.

4. AN ALTERNATIVE EXPLANATION OF THE C-VALUE PARADOX

From these hypotheses of *artificial* genome growth I propose a possible explanation for the C-value Paradox in real biology. The explanation is inspired by the removal bias and modification point depth models. The general form is very simple: genomes grow because insertion of material is more positively correlated (or less negatively correlated!) with fitness improvement than is deletion of material.[8] This notion is relatively intuitive given the existing EC bloat hypotheses, though its application to the C-value Paradox has not, to my knowledge, been stated.

Here is one concrete example of how this principle might be applied to DNA. Consider the process, either via bulk modification, transposition, or other method, whereby a strand of DNA is either snipped out, transposed (snipped and reinserted), or copied. Presume that these processes are among the machinery by which DNA undergoes genetic mutation and other adaptive change. Presume also that elimination of genes in the genome on average tends to be damaging to the individual. Then we ask the following question: which process (deletion, transposition, or copying) is more likely to destroy genes? Here they are in order of increasing damage.

- *Copying* may damage those genes which presently straddle or otherwise rely on material at the chosen insertion point.
- *Transposition* damages the same genes as copying. It also damages genes straddling or otherwise relying on the two end-points of the sliced-out material. This is effectively three times as many gene locations (or perhaps two if the end-points are close together).
- *Deletion* damages not only the genes straddling the two end-points of the sliced-out material, but of course also eliminates any genes located within the sliced-out region.

Importantly this order is also the same as: material added; no change; material removed. Thus the application of these processes constitute a kind of "removal bias" applied to the genome. Increase in DNA material may be explained simply because it is less likely to destroy the individual.

Unlike the Nucleoskeletal and Nucleotypic hypotheses, the proposed explanation provides an actual functional mechanism by which genome growth may occur, rather than simply assuming that genome growth is the *only* way to fulfill the needs of larger nuclei, larger and slower-dividing cell bodies, or other cellular parameters. And unlike the Junk DNA and Selfish DNA hypotheses, the hypothesis presented here expressly permits processes of material deletion, and in fact benefits from them (though it does not formally require them). It is worthwhile noting that examples of significant DNA deletion (for example, in *Arabidopsis* [9]) are not necessarily evidence against the theory: the presence of a potent DNA reduction force does not preclude a more general, possibly weaker, DNA-increasing force. Finally, like the Junk DNA hypothesis, the proposed hypothesis also suggests a linear or sublinear increase in DNA size.

[8]Credit is due Kenneth De Jong, who first conveyed the general concept to me in conversation, though in the context of code bloat in evolutionary computation.

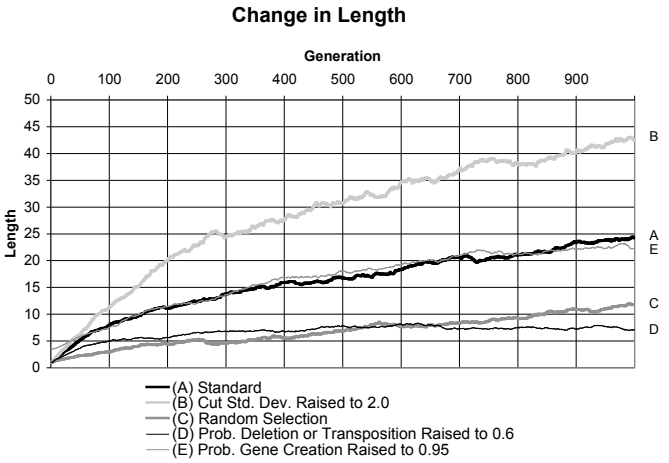

Figure 1: Change in length of the genome with increasing generations. Shown are the standard proposed values and four variations.

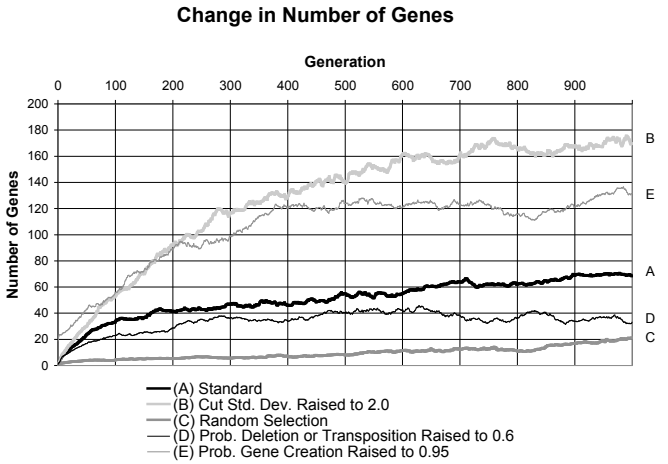

Figure 2: Change in number of genes with increasing generations. Shown are the standard proposed values and four variations.

Some caveats. This explanation is not derived from observation of biological data: it is instead inspired by observation of a similar phenomenon in an artificial system. This, I believe, is the likely reason why the theory has not been (to my knowledge) considered by the biological literature despite its obviousness and simplicity. But though code bloat and the C-value Paradox bear a great many similarities, this could be *entirely coincidental.* As such, it is possible that this hypothesis will not stand the test of biological experiment, and further, I am not qualified to make such an assessment. Instead, I offer this hypothesis and argue for it based on its novelty and straightforwardness.

The explanation is also not intended to replace the existing hypotheses. Transposable element evidence is supportive of the other hypotheses, and strand slippage seems to be well supported. There are problems with these hypotheses, but they are hardly ruled out by these difficulties. Instead, I view this new hypothesis as a complement to the others. It is certainly possible for the C-value Paradox to be caused by a variety of growth factors.

5. AN ILLUSTRATIVE EXAMPLE

The example employs a basic generational EC algorithm using a population of 100 individuals, run up to 1000 generations, using tournament selection of size 2 and no elitism. Each individual in the population contains a single chromosome, which is represented by a real-valued half-open interval $[0, c)$ where c is the length, or C-value, of the chromosome. Along this interval lie "genes", which are simply half-open intervals $[x, y)$ representing the gene region along the chromosome, $0 \leq x < y \leq c$. Therefore, $y - x$ is the (non-zero) length of the gene. Individuals' chromosomes in the first generation are each created with an initial length of 1.0. For each chromosome, we flip a true/false coin of probability 0.5 until it comes up false. The number of trues is the number of genes added to that chromosome. Genes are created with a random length uniformly chosen from 0 to 0.01 and a random location chosen uniformly from 0 to 1.0. Overlaps are not permitted. If a gene extends beyond the boundary of the genome or straddles an existing gene, it is replaced with another randomly-generated gene.

Breeding is asexual. During breeding an individual may be subjected to some number of replication events (these events being deletion, copying, and transposition). The number of events is defined as before by counting the number of flips of a true/false coin of probability 0.5 until it comes up false. Events are formed in the following way. First, a copy start-location is chosen at random in the chromosome $[0, c)$. Second, a copy end-location is chosen at random from a normal distribution with the start-location as mean and 1.0 as the standard deviation. If the end-location is outside the bounds of the chromosome, a new start-location and end-location are again chosen, and so on, until a valid end-location is formed. We then copy into a buffer the interval between the start- and end-locations, including all genes falling in this interval but not straddling the start- and end-locations.

Next we flip a coin of probability 0.5 to determine whether or not to delete the copy-interval. If so, the interval is sliced out of the chromosome, shortening it, and removing all genes which fell in the interval or straddled its endpoints. Finally, we flip a coin of probability 0.5 to determine whether or not to reinsert the buffer into the chromosome. If so, we select a random insertion location, delete from the chromosome all genes which straddle that location, and then insert the buffer at that location. Thus there is a 0.25 chance each that the event may delete, copy, transpose, or do nothing.

There is a problematic sink condition when no selection is involved: if a chromosome has zero genes, it cannot increase in number of genes. To remedy this, if during deletion a chromosome is emptied of genes, it must reinsert the buffer. If all genes have been killed during insertion or deletion, one new gene is added to the genome. Approximately half of the chromosomes will start with zero genes in the first generation, but very rapidly nearly all chromosomes have at least one gene.

This is all a very crude arrangement. Among other things, it presumes that the chromosome has a real-valued length; it has completely preposterous values for typical gene length,

probability of replication events, and length of the sliced-out region; it presumes that insertion points are completely independent of deletion regions; and most problematically, it assumes that the number of events is independent of the length of the chromosome. This is freely admitted, as the intention of the example is only to provide an illustration of the general idea of how the hypothesis might operate, not a proof of its correctness. For several of the decisions (such as restriction of sliced-out region length and the independence of event probability from the length of the chromosome) I chose to err on the conservative side in the sense that those decisions would be expected to help the hypothesis *less*.

We are left with the definition of fitness. This is not an easy task: modeling fitness in a real biological system is nontrivial. Here again I will use an extremely conservative and crude measure of fitness for our example: an individual's fitness is the number of genes, or some value $M = 5$, whichever is smaller. Larger fitnesses are preferred. The intuition here is that there is *some* fitness advantage to not deleting genes, but it's not strong: having many more than than M genes conveys no advantage over having just a few more than M genes.

5.1 Results

I performed five variations on this setup to gauge genome growth. Each experiment ran for 50 independent trials using the ECJ evolutionary computation system [14]. Figures 1 and 2 show the mean growth in genome length and in number of genes for the five variations respectively. Comparisons among the experiments applied an ANOVA at $p = 0.05$ on the log of the generation-1000 value of the number of genes and of the genome lengths (logs were used to transform strongly right-tailed results to a more normal distribution).

Standard. The initial experiment was performed using exactly the parameters described earlier. Recall that the conservative fitness measure is simply the minimum of 5 and the number of genes in the genome. Thus having a very large number of genes *should* have a diminishing return in fitness advantage, and so one would expect genome growth to taper off rapidly. Number of genes does begin to taper, but the average length increases almost linearly. By 1000 generations, these values achieved a surprisingly large mean length of 68.32 and a mean number of genes of 24.18.

Changing the Cut Size. What happens when the average cut size is changed? This affects the ratio of cut size to typical gene length and also the ratio of cut size to typical genome length. I modified the cut procedure to use a standard deviation of 2.0 for choosing cut sizes rather than a standard deviation of 1.0. As can be seen, the result is a 3x jump in number of genes and a 2x jump in length. Again, growth in number of genes begins to taper, but growth in length is very nearly linear. Due to large variances and the transformation via logarithm, the difference in number of genes was barely not statistically significant, but the increase in length was significant. In experiments (not shown here) involving standard deviations of 10 and 20, the number of genes and genome length grew even further.

Random Selection. Was the result due to a random walk? The number of genes and genome length both start out small, and so a random walk would be expected to increase them slowly because they are both bounded below at 0. But though the walk did head away from the lower bound, the effect was rather small. With random selection, the mean length only reached 12, and the mean number of genes only reached 21. The difference with the "standard" experiment was highly statistically significant.

Sensitivity to Deletion or Transposition. Deletion, Copying, and Transposition all had the same probability of occurrence. One would expect that less copying would result in slower growth. To test this, I changed the cut-probability to 0.6, resulting in a major dampening of growth. Growth occurred until at some point the forces of selection and additional deletion converged. Change in length was not statistically significantly different from random selection: but change in genes was statistically significantly larger than random selection. The difference between this and the "standard" experiment was statistically significant.

Initial Seeding. Because the coin-flip probability for creating genes was set to 0.5, many individuals would have very few, or even 0, initial genes. This gave individuals with large chromosomes a significant and perhaps unfair advantage, resulting in a linear increase to many genes and a very long genome. What if populations were seeded with many genes to start with? I changed the probability of gene creation to 0.95, which raised the average number of initial genes to approximately 19 (from 5). Accordingly, I increased the initial genome length to 3.5 to maintain approximately the same gene density. Surprisingly, the result produced nearly identical change in length as the "standard" experiment, but with twice as many genes (statistically significant).

Even with this very conservative measure of fitness, most variations on the experiment produced longer genomes with many more genes than random selection. Given the intuitive nature of the hypothesis, this is hardly surprising, though it *is* surprising that the genome length tends to grow seeming without bound, even if the number of genes tends to taper off. This growth is considerably higher than the natural growth due to the random walk.

Of course, real biological fitness is different from this: no organism has only 5 genes; and it is likely that destruction of any one of a large number of genes in the organism could put it at selective disadvantage, if not kill it outright. Sprouting multiple copies of the same gene would also help defend against damage due to DNA deletion (shades of the Protection Against Recombination hypothesis). Still, these differences all would seem to point to even stronger genome growth than in the conservative illustration put here.

6. CONCLUSIONS

This paper presented a new hypothesis explaining the phenomenon of accumulation of non-coding DNA in genomes in biology. This phenomenon forms a central question of the so-called C-value Paradox. But the germ of the hypothesis did not come from biological experiment but rather from similar phenomena found in artificial genotype representations in evolutionary computation. Representations such as genetic programming program trees, arbitrary-length lists of machine-language instructions, rule sets, etc., can all suffer from the scourge of code bloat, where the genome grows without a reasonably justifying improvement in fitness.

The hypothesis suggests at the high level that one factor in genome growth may be a selective bias towards events which increase the amount of DNA rather than decrease it. In one concrete example, selection may prefer DNA-copying events over transposition events, and transposition events over deletion events, because DNA copying is less likely to destroy important gene material than transposition is, and transposition is in turn less likely to do so than deletion is. Unlike the Junk DNA and Selfish DNA hypotheses, this new hypothesis relies on natural selection itself as the driving force, rather than resorting to drift or "selfish" gene-level selection. And unlike the Nucleoskeletal and Nucleotypic hypotheses, the new hypothesis provides an actual functional explanation for growth.

I also presented a simple illustration of the process occurring in simulation as a proof of concept. The simulation example is admittedly extremely crude, but even with the conservative assumptions it makes, the phenomenon is still present. While the number of genes tends to reach an upper limit (the predicted outcome of the simulation), the length of the genome seems to grow linearly and without bound.

I do not suggest that this hypothesis should replace the others: in fact I believe that much of the C-value Paradox may rightly be explained by them. Instead I suggest this hypothesis may describe a process which is at least partially responsible for the phenomenon. The hypothesis has not yet been tested against biological data: but the straightforwardness and simplicity of the hypothesis recommend it for consideration by the theoretical biology community. Evolutionary computation steals a lot from biological theory. It's not often the field gets to give something back.

7. ACKNOWLEDGMENTS

Thanks to Sanjeev Kumar for his considerable input, to Stephen Mount for sanity-checking the work, and to the reviewers for their highly constructive comments. Thanks also to Jeff Bassett, Liviu Panait, Gabriel Balan, Kenneth De Jong, and Dana Richards.

8. REFERENCES

[1] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone. *Genetic Programming – An Introduction; On the Automatic Evolution of Computer Programs and its Applications*. Morgan Kaufmann, dpunkt.verlag, Jan. 1998.

[2] M. D. Bennett. Nuclear DNA content and minimum generation time in herbaceous plants. *Proceedings of the Royal Society of London, Series B*, 181:109–135, 1972.

[3] T. Blickle. *Theory of Evolutionary Algorithms and Application to System Synthesis*. PhD thesis, Swiss Federal Institute of Technology, Zurich, Nov. 1996.

[4] T. A. Brown. *Genetics: A Modern Approach*. Chapman and Hall, London, 1992.

[5] T. Cavalier-Smith. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate and the solution of the DNA C-value paradox. *Journal of Cell Science*, 34:247–278, 1978.

[6] T. Cavalier-Smith. How selfish is DNA? *Nature*, 285:617–618, June 1980.

[7] B. Charlesworth, P. Sniegowski, and W. Stephan. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371:215–220, September 1994.

[8] R. Dawkins. *The Selfish Gene*. Oxford University Press, new edition, 1989.

[9] K. M. Devos, J. K. M. Brown, and J. L. Bennetzen. Genome size reduction through illegitimate recombination counteracts genome expansion in *arabidopsis*. *Genome Research*, 12(7):1075–1079, 2002.

[10] W. F. Doolittle and C. Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284:601–603, April 1980.

[11] T. R. Gregory. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews*, 76:65–101, 2001.

[12] W. B. Langdon. Quadratic bloat in genetic programming. In D. Whitley, D. Goldberg, E. Cantu-Paz, L. Spector, I. Parmee, and H.-G. Beyer, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, pages 451–458, Las Vegas, Nevada, USA, 10-12 July 2000. Morgan Kaufmann.

[13] S. Luke. Modification point depth and genome growth in genetic programming. *Evolutionary Computation*, 11(1):67–106, 2003.

[14] S. Luke. ECJ 12: An evolutionary computation research system in Java. Available at http://cs.gmu.edu/ eclab/projects/ecj/, 2004.

[15] P. Nordin and W. Banzhaf. Complexity compression and evolution. In L. Eshelman, editor, *Genetic Algorithms: Proceedings of the Sixth International Conference (ICGA95)*, pages 310–317, Pittsburgh, PA, USA, 15-19 July 1995. Morgan Kaufmann.

[16] S. Ohno. *Evolution by Gene Duplication*. Springer, New York, 1970.

[17] L. E. Orgel and F. H. C. Crick. Selfish DNA: the ultimate parasite. *Nature*, 284:604–607, April 1980.

[18] M. Pagel and R. A. Johnstone. Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proceedings: Biological Sciences*, 249(1325):119–124, August 1992.

[19] A. Prachhumwat, L. DeVincentis, and M. F. Palopoli. Intron size correlates positively with recombination rate in *caenorhabditis elegans*. *Genome Research*, 166:1585–1590, March 2004.

[20] S. F. Smith. *A Learning System Based on Genetic Adaptive Algorithms*. PhD thesis, Computer Science Department, University of Pittsburgh, 1980.

[21] T. Soule and J. A. Foster. Removal bias: a new cause of code growth in tree based evolutionary programming. In *1998 IEEE International Conference on Evolutionary Computation*, pages 781–186, Anchorage, Alaska, USA, 5-9 May 1998. IEEE Press.

[22] W. A. Tackett. *Recombination, Selection, and the Genetic Construction of Computer Programs*. PhD thesis, University of Southern California, Department of Electrical Engineering Systems, USA, 1994.

[23] C. A. Thomas. The genetic organization of chromosomes. *Annual Review of Genetics*, (5):237–256, 1971.